A Language-independent Sense Clustering Approach for Enhanced WSD

Michael Matuschek[*], Tristan Miller[*], and Iryna Gurevych[*†]

[*]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
http://www.ukp.tu-darmstadt.de/
[†]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
http://www.dipf.de/

Abstract

We present a method for clustering word senses of a lexical-semantic resource by mapping them to those of another sense inventory. This is a promising way of reducing polysemy in sense inventories and consequently improving word sense disambiguation performance. In contrast to previous approaches, we use Dijkstra-WSA, a parameterizable alignment algorithm which is largely resource- and language-agnostic. To demonstrate this, we apply our technique to GermaNet, the German equivalent to WordNet. The GermaNet sense clusterings we induce through alignments to various collaboratively constructed resources achieve a significant boost in accuracy, even though our method is far less complex and less dependent on language-specific knowledge than past approaches.

1 Introduction

Lexical-semantic resources (LSRs) are a prerequisite for many key natural language processing tasks. However, it is nowadays widely recognized that not every resource is equally well suited for each task. For word sense disambiguation (WSD), which is the focus in this paper, the Princeton WordNet (Fellbaum, 1998) is the predominant sense inventory for English because of its free availability, its comprehensiveness, and its use in dozens of previous studies and data sets. For German, GermaNet (Hamp and Feldweg, 1997) is the German equivalent to WordNet and has positioned itself as the reference resource for WSD, although systematic investigation of German WSD has only recently begun (Broscheit et al., 2010; Henrich and Hinrichs, 2012).

There is much evidence to suggest that the sense distinctions of expert-built wordnets are far subtler than what is typically necessary for real-world NLP applications, and sometimes even too subtle for human annotators to consistently recognize. This point has been made specifically for WordNet (Ide and Wilks, 2006), but is just as applicable to other expert-built resources (Jorgensen, 1990). This makes improving upon experimental results difficult, while at the same time the downstream benefits of improving WSD on these LSRs are often not clearly visible.

Using a different sense inventory could solve the problems inherent to expert-built LSRs, and recently collaboratively constructed resources, such as Wiktionary and Wikipedia, have been suggested (Mihalcea, 2007). These resources are attractive because they are large, freely available in many languages, and under continuous improvement. However, they still contain considerable gaps in coverage, few large-scale sense-annotated corpora use them, and for some word categories their senses are also rather fine-grained. Much prior work has therefore focused instead on enhancing wordnets by decreasing their granularity through (semi-)automatic clustering of their senses. However, until now, the focus of attention has almost exclusively been the English WordNet. While it has been shown that such clustering significantly enhances both human interannotator agreement (Palmer et al., 2007) and automatic WSD performance (Snow

et al., 2007), the previous approaches had been specifically tailored towards this resource, making the applicability to other LSRs, let alone other languages, difficult.

In this paper, we describe a new, fully automated approach to the granularity problem which taps the benefits of collaboratively constructed LSRs without the drawbacks of using them as wholesale replacements for other LSRs. Specifically, we induce a clustering of a resource's senses by first mapping them to those in the other resources, and then grouping source senses which map to the same target sense. This results in a coarse-grained sense inventory. In contrast to previous alignment-based clustering techniques, we use Dijkstra-WSA, a state-of-the-art sense alignment algorithm which is highly parameterizable as well as resource- and language-agnostic. This allows us to produce clusterings based on several different German resource alignments, for which we conduct in-depth analyses and evaluations. To demonstrate the language-independence of our approach, we produce clusters for both GermaNet and WordNet, though our algorithm is easily applicable to many resource pairs.

2 Related work

Clustering fine-grained sense distinctions into coarser units has been a perennial topic in WSD. Past approaches have included using text- and metadata-based heuristics to derive similarity scores for sense pairs in electronic dictionaries (Dolan, 1994; Chen and Chang, 1998), exploiting semantic hierarchies to group senses by proximity or ancestry (Peters et al., 1998; Buitelaar, 2000; Mihalcea and Moldovan, 2001; Tomuro, 2001; Ide, 2006), grouping senses which lexicalize identically when manually translated (Resnik and Yarowsky, 2000), using distributional similarity of senses (Agirre and Lopez de Lacalle, 2003; McCarthy, 2006), exploiting disagreements between human annotators of sense-tagged data (Chklovski and Mihalcea, 2003), heuristically mapping senses to learned semantic classes (Kohomban and Lee, 2005), and deep analysis of syntactic patterns and predicate–argument structures (Palmer et al., 2004; Palmer et al., 2007).

Comparison of these approaches is hampered by the fact that evaluations often are not provided in the papers, are applicable only for the particular LSR used in the experiment, do not provide a random baseline for reference, and/or provide only intrinsic measures such as "reduction in average polysemy" which do not directly speak to the clusterings' correctness or utility for a particular task. Though many of the above authors cite improved WSD as a motivation for the work, most of them do not actually investigate how their clusterings impact state-of-the-art disambiguation systems. The only exception is Palmer et al. (2007), who compare results of a state-of-the-art WSD system, as well as human interannotator agreement, on both fine-grained and clustered senses. To ensure that the measured improvement was not due solely to the reduced number of sense choices for each word, they also evaluate a random clustering of the same granularity.

Apart from the above-noted approaches, there has also been interest recently in techniques which reduce WordNet's sense granularity by aligning it to another, more coarse-grained resource at the level of word senses. Navigli (2006) induces a sense mapping between WordNet and the *Oxford Dictionary of English* (Soanes and Stevenson, 2003) on the basis of lexical overlaps and semantic relationships between pairs of sense glosses. WordNet senses which align to the same *Oxford* sense are clustered together. The evaluation is similar to that later used by Palmer et al. (2007), except that rather than actually running a WSD algorithm, Navigli expediently takes the raw results of a Senseval WSD competition (Snyder and Palmer, 2004) and does a coarse-grained rescoring of them. The improvement in accuracy is reported relative to that of a random clustering, though unlike in Palmer et al. (2007) there is no indication that the granularity of the random clusters was controlled. It is therefore hard to say whether the clustering really had any benefit.

Snow et al. (2007) and Bhagwani et al. (2013) extend Navigli's approach by training machine learning classifiers to decide whether two senses should be merged. They make use of a variety of features derived from WordNet as well as external sources, such as the aforementioned *Oxford*–WordNet mapping. They also improve upon Navigli's evaluation technique in two important ways: first, they ensure their baseline random

clustering has the same granularity as their induced clustering, and second, the random clustering performance is computed precisely rather than estimated stochastically. While their methods result in an improvement over their baseline, they do require a fair amount of annotated training data, and their features are largely tailored towards WordNet-specific information types. This makes the methods' transferability to resources lacking this information rather difficult.

In this paper, we go beyond this previous work in two ways. First, we employ Dijkstra-WSA (Matuschek and Gurevych, 2013), a state-of-the-art alignment algorithm with the attractive property of being largely resource- and even language-agnostic. This makes the alignment (and hence, the clustering approach) easily applicable to many different resource combinations, though we expect its performance to be competitive with far more complex and resource-specific approaches.

Second, thanks to the flexibility of Dijkstra-WSA, we can perform a deeper comparative analysis of alignment-based clusterings against not one but three different LSRs. We investigate how the different properties of these resources influence the alignments and clusterings, particularly with respect to accuracy across parts of speech. This is the first time such a detailed analysis is presented. We focus on collaboratively constructed LSRs, as their emergence has led to an ongoing discussion about their quality and usefulness (Zesch et al., 2007; Meyer and Gurevych, 2012; Krizhanovsky, 2012; Gurevych and Kim, 2012; Hovy et al., 2013). Our work aims to contribute to this discussion by investigating the crucial aspects of granularity and coverage.

3 Alignment-based clustering

3.1 Task description

Word sense clustering is the process, be it manual or automatic, of identifying senses in an LSR which are similar to the extent that they could be considered the same, slight variants of each other, or perhaps subsenses of the same broader sense. Its purpose is to merge these senses (i.e., to consider the set of clustered senses as a single new sense) so as to facilitate usage of the sense inventory in applications which benefit from a lower degree of polysemy, such as machine translation, where lexical ambiguity is often preserved across certain language pairs, making fine-grained disambiguation superfluous. For example, the two WordNet senses of *ruin*—"destroy completely; damage irreparably" and "reduce to ruins"—are very closely related and could be used interchangeably in many contexts.

One way to achieve such a clustering is *word sense alignment* (WSA), or *alignment* for short. An alignment is formally defined as a list of pairs of senses from two LSRs, where the members of each pair represent the same meaning. When it is not restricted to 1:1 alignments, it is possible that a sense s in one LSR A is assigned to several senses $t_1, \ldots, t_n$ in another LSR B. Assuming that all alignments are correct, this implies that $s \in A$ is more coarse-grained and subsumes the other senses, which in turn can be considered as a sense cluster within B. For example, the aforementioned senses of *ruin* could both be aligned to the Wiktionary sense "to destroy or make something no longer usable" and thereby clustered.

3.2 Lexical-semantic resources

For our experiments we align GermaNet, a German wordnet, to three different collaboratively constructed German LSRs: Wikipedia, Wiktionary, and OmegaWiki. Our goal is to demonstrate that effective sense clustering is possible for resources in languages other than English using a language-agnostic alignment approach.

Moreover, we aim to cover two popular dictionary resources which are at different stages of development regarding size and coverage (OmegaWiki and Wiktionary) as well as the most popular collaboratively constructed encyclopedia (Wikipedia), which was not designed as a lexicographic knowledge source but is widely used in NLP nonetheless (Zesch et al., 2007; Milne and Witten, 2008). As the detailed results of the alignment are of secondary interest here (being exhaustively discussed in Matuschek and Gurevych (2013)), we focus on a discussion of the clusterings which are derived from the alignment and relate these results to the properties of the LSRs involved. For convenient usage in our clustering framework, we use the LSR versions found in the unified resource UBY (Gurevych et al., 2012).

GermaNet (Hamp and Feldweg, 1997) is an expert-built computational lexicon for German and thus the counterpart to WordNet. It is organized into synsets (over 84 500 in version 8.0, which we use) connected via semantic relations.

Wikipedia is a free, multilingual, collaboratively written online encyclopedia and one of the largest publicly available knowledge sources. Each article usually describes a distinct concept which is connected to other articles by means of hyperlinks. UBY contains a snapshot of the German edition from 16 August 2009 with around 834 000 articles.

Wiktionary is a dictionary "sister project" of Wikipedia. For each word, multiple senses can be encoded, and these are usually also represented by glosses. There are also hyperlinks which lead to synonyms, hypernyms, meronyms, etc. UBY's 6 April 2011 snapshot of the German edition contains around 72 000 entries.

OmegaWiki is another freely editable online dictionary. Unlike in Wikitionary, there are no distinct language editions; OmegaWiki is comprised of language-independent concepts ("defined meanings") which bear lexicalizations in various languages. These are connected by semantic relations as in WordNet. UBY uses a database dump from 3 January 2010, which contains slightly less than 47 000 concepts and lexicalizations in over 470 languages.

3.3 Dijkstra-WSA

Dijkstra-WSA is the graph-based word sense alignment algorithm which we use to infer the clusterings. It consists of three steps: (i) the initial construction of the graphs, (ii) the identification of valid alignments using a shortest path algorithm, and (iii) an optional similarity-based backoff for senses which could not be aligned.

Graph construction. The set of senses (or synsets, if applicable) of an LSR is represented as a set of nodes V where the set of edges $E \subseteq V \times V$ between these nodes represents semantic relatedness between them. This is called a *resource graph*. For deriving the edges, one can use semantic relations (such as hyponymy), hyperlinks (for Wikipedia), or other relatedness indicators provided by the resource. For sparse LSRs such as Wiktionary, it is a viable option to increase the density by adding edges between senses s_1 and s_2 if a monosemous term t with sense s_2 is included in the gloss of s_1. For example, one can link a sense of *Java* to *programming language* if the latter term is included in the former's definition text. This so-called *linking of monosemous lexemes* proved to significantly enhance the graph density (and hence, the recall of the alignment) with only a minor loss in precision.

Computing sense alignments. For the two resource graphs A and B, edges representing trivial alignments are introduced first. Alignments are trivial if two senses have the same attached lexeme in A and B and this lexeme is also monosemous in each resource. For example, if the noun phrase *programming language* is contained in either resource and has exactly one sense in each one, we can directly infer the alignment.

Next, we consider each still unaligned sense $s \in A$. We first retrieve the set of target senses $T \subset B$ with matching lemma and part of speech (e.g., *Java (island)* and *Java (programming language)*) and compute the shortest path to each of them with Dijkstra's shortest path algorithm (Dijkstra, 1959). The candidates in T with a distance below a certain threshold (estimated on a development set considering the graph size and density) are selected as alignment targets, and the algorithm continues until either all senses are aligned or no path can be found for the remaining senses. The intuition behind this is that the trivial alignments serve as "bridges" between A and B, such that a path starting from a sense s_1 in A traverses edges to find a nearby already aligned sense s_2, "jumps" to B using a cross-resource edge leading to t_2 and then ideally finds an appropriate target sense t_1 in the vicinity of t_2. In this example, the bridge *programming language* would enable the correct identification of two equivalent senses of *Java*. Note that our definition allows computation of one-to-many alignments, which are a prerequisite for the subsequent clustering step we describe in Section 3.1. Also note that with each successful alignment, edges are added to the graph so that a different ordering of the considered senses leads to different results; these differences were in no case statistically significant, however.

Similarity-based backoff. Alignments found by Dijkstra-WSA are complementary to those usually found by text similarity–based approaches. We therefore use a hybrid approach which first uses Dijkstra-WSA and falls back to gloss similarity for those cases where no target could be found in the graph. This significantly increases the alignment recall, so in order to better understand the consequences for our clustering system, we run Dijkstra-WSA both with and without this backoff. However, we do not employ a machine learning component; to keep the approach as knowledge-poor as possible, we follow the approach by Henrich et al. (2011) and align to the candidate with the greatest similarity.

4 Evaluation

4.1 Methodology

A common extrinsic method for evaluating sense clusterings is to take the raw assignments made by existing word sense disambiguation systems on a standard data set and then rescore them according to the clustering. That is, a system is considered to have correctly disambiguated a term not only if it chose a correct sense specified by the data set's answer key, but also if it chose any other sense in the same cluster as a correct one. Of course, any clustering whatsoever is likely to increase accuracy, simply by virtue of there being fewer answers for the system to choose among. To account for this, accuracy obtained with each clustering must be measured relative to that of a random clustering of equivalent granularity.[1]

The random clustering score for each instance in the data set can be determined mathematically. Snow et al. (2007) and Bhagwani et al. (2013) use

$$\sum_{c \in C} \frac{|c|\,(|c|-1)}{N\,(N-1)}, \tag{1}$$

where C is the set of clusters over the N senses of a given term, and $|c|$ is the number of senses in the cluster c. However, this formula is accurate only when the gold standard specifies a single correct answer for the instance. In practice, WSD data sets can specify multiple possible correct senses for an instance, and a system is considered to have correctly disambiguated the target if it selected any one of these senses. The Senseval-3 all-words corpus used by Snow et al. (2007) and Bhagwani et al. (2013) is such a data set (some 3.3% of the instances have two or more "correct" senses) so the scores they report underestimate the accuracy of the random baseline and inflate their clustering methods' reported improvement.

To arrive at a formula which works in the general case, consider that for an instance where the target word has N senses, g of which are correct in the given context, and one of which is an incorrectly chosen sense, the total number of ways of distributing these senses among the clusters is

$$N \cdot \binom{N-1}{g} = \frac{N!}{g!(N-g-1)!}. \tag{2}$$

Of these, the number of distributions which cluster the incorrectly chosen sense together with none of the correct senses is

$$\sum_{c \in C} |c| \binom{N-|c|}{g} = \sum_{c \in C} \frac{|c|\,(N-|c|)!}{g!\,(N-|c|-g)!}, \tag{3}$$

where the summation includes only those clusters where $N - |c| \geq g$. The probability that the incorrectly chosen sense is clustered together with at least one correct sense is therefore

$$1 - \sum_{c \in C} \frac{|c|\,(N-|c|)!\,(N-g-1)!}{N!\,(N-|c|-g)!} \tag{4}$$

or, recast for ease of programmatic computation,

$$1 - \sum_{c \in C} \frac{|c| \prod_{i=0}^{g-1} (N-|c|-i)}{\prod_{i=0}^{g} (N-i)}. \tag{5}$$

For the case where there really is only one correct gold-standard answer, Formula 4 becomes

$$\begin{aligned} 1 - \sum_{c \in C} \frac{|c|\,(N-|c|)}{N\,(N-1)} &= \sum_{c \in C} \frac{|c|}{N} - \sum_{c \in C} \frac{|c|\,(N-|c|)}{N\,(N-1)} \\ &= \sum_{c \in C} \frac{|c|\,(|c|-1)}{N\,(N-1)}, \end{aligned} \tag{6}$$

which agrees with Formula 1 above.

To compute the clustered scoring, including that of the random clusterings, we use the free DKPro WSD framework (Miller et al., 2013).

[1] Controlling for granularity is vital, since it is trivial to construct clusterings which effect arbitrarily high WSD accuracy. Consider the extreme case where for each word, *all* the senses are clustered together; this clustering would have 100% WSD accuracy and thus easily beat an uncontrolled random baseline, but not a granularity-controlled one.

	aff.	imp.	%
OmegaWiki (DWSA)	438	130	29.7
OmegaWiki (sim. only)	712	165	23.2
OmegaWiki (w/backoff)	872	205	23.5
Wiktionary (DWSA)	1355	311	23.0
Wiktionary (sim. only)	1463	349	23.8
Wiktionary (w/backoff)	1797	349	19.4
Wikipedia (DWSA)	773	120	15.5
Wikipedia (sim. only)	710	158	22.2
Wikipedia (w/backoff)	852	147	17.3

Table 1: Number and percentage of lexical items in the data set affected and improved by the clusterings. The slight proportional decrease in improved items in some configurations results from an improved alignment recall using the backoff.

4.2 Data sets and algorithms

To our knowledge, there are currently only two German-language sense-annotated corpora, both of the "lexical sample" variety: DeWSD (Broscheit et al., 2010) and WebCAGe (Henrich et al., 2012). At the time of writing only the latter was available to us, and so is the one used in our study. With 10 429 instances of 2719 lexical items annotated with GermaNet 8.0 senses, WebCAGe 2.0 is significantly larger and more up to date than DeWSD, which has 1154 instances of 40 lexical items annotated with GermaNet 5.1 senses. As with the Senseval-3 data set, many WebCAGe instances specify multiple gold-standard senses.

German-language WSD is still in its infancy; the only results reported so far on WebCAGe are for various weakly supervised, Lesk-like systems (Henrich and Hinrichs, 2012).[2] For our extrinsic cluster evaluation, we therefore rescore the sense assignments made by their *lsk_Ggw+Lgw* system, the best-performing system (in terms of recall and F_1) when run on the entire WebCAGe 2.0 corpus.

4.3 Experiments on GermaNet

GermaNet–OmegaWiki. When only Dijkstra-WSA is used for clustering, the clusters are small and few in number. This results in few lexical items in the data set being affected by the clustering, and is in line with the observation made in Matuschek and Gurevych (2013) that graph-based alignments usually yield good precision at the expense of recall. So although relatively few senses are aligned and subsequently clustered, the clusters seem mostly correct, which is indicated by the significant overall improvement. The first line of Table 1 shows how many of the 10 429 instances of the evaluation data set were actually affected by this clustering configuration, and of these how many saw an increase in accuracy over the random baseline (which is an indicator of the validity of the clusters).

For adjectives (the smallest part-of-speech group in the data set) there is almost no clustering at all, as for most senses Dijkstra-WSA identified no targets, or only one target. The situation was better for nouns and verbs; while the clusters are not large (usually 2–3 senses), the high-precision clustering did improve the results. Nouns especially saw a statistically significant[3] improvement over the random clustering (1.6 percentage points). The upper third of Table 2 shows the full results for this setup. The table shows the original accuracy score without clustering (*none*), the accuracy with our clustering (*WSA*), the accuracy with random clustering of equivalent granularity (*rand.*), and the difference between the latter two ($\pm$).

When gloss similarity is used in isolation, we achieve a higher alignment recall and thus larger clusters; this way, we are able to cluster a substantial number of adjectives, leading to an increase in WSD performance. However, the overall results are worse due to the lower precision for nouns.

When we employ the backoff to improve the recall of the graph-based alignment (i.e., a combination of both approaches), we get more and larger clusters (see third line of Table 1), leading to a significant improvement in WSD accuracy for nouns and verbs (Table 2). Although alignment precision for this setup was reported to be generally worse than for Dijkstra-WSA alone, the alignments are seemingly still precise enough to form meaningful clusters with only a few errors.

A good example is the verb *markieren* ("to mark"), whose only sense in OmegaWiki ("somehow tag for later reference") is aligned to two

[2] Broscheit et al. (2010) evaluate a graph-based WSD system, albeit only on the DeWSD corpus.

[3] All significance statements in this paper are based on McNemar's test at a confidence level of 5%.

			OmegaWiki			Wiktionary			Wikipedia		
		none	**rand.**	**WSA**	±	**rand.**	**WSA**	±	**rand.**	**WSA**	±
no backoff	noun	51.1	60.9	62.5	1.6*	75.1	77.2	**2.1***	75.1	76.3	1.2*
	verb	43.1	45.8	46.6	0.8*	60.1	61.8	**1.7***	—	—	—
	adj.	43.3	45.0	45.0	0.0	82.5	83.0	0.5	—	—	—
	all	48.1	55.3	56.5	1.2*	71.2	73.0	**1.8***	—	—	—
sim. only	noun	51.1	61.6	62.7	1.1*	72.3	73.8	1.4*	70.5	71.6	1.1*
	verb	43.1	55.5	56.3	0.8*	58.7	58.7	0.0	—	—	—
	adj.	43.3	61.6	62.1	0.5	65.9	66.3	0.4	—	—	—
	all	48.1	59.8	60.7	0.9*	67.8	68.7	0.9*	—	—	—
w/backoff	noun	51.1	66.9	68.5	1.6*	83.2	85.3	**2.1***	76.6	78.6	2.0*
	verb	43.1	56.0	57.3	1.3*	73.7	74.3	0.6	—	—	—
	adj.	43.3	61.1	62.0	**0.9**	87.9	87.8	−0.1	—	—	—
	all	48.1	63.3	64.7	1.4*	80.7	82.2	1.5*	—	—	—

Table 2: WSD accuracy (F-score) by POS, using clusterings derived from alignments of GermaNet to various resources, via Dijkstra-WSA without (top) and with (bottom) the similarity-based backoff, or via gloss similarity only (middle). Boldface marks best results per POS; asterisks mark statistically significant differences from the granularity-controlled random baseline.

GermaNet senses, one each for text and territorial marking. The difference in polysemy between GermaNet and OmegaWiki (see Table 3) pays off here, as the coarse OmegaWiki sense subsumes the GermaNet senses. This is exactly the intended effect when this kind of clustering is performed.

However, there are also many notable gaps in coverage (Table 3)—even some commonly used terms are missing from OmegaWiki altogether, leaving their GermaNet senses unaligned and unclustered. This underrepresentation of lemmas and senses can be attributed to the fact that OmegaWiki, in comparison to Wiktionary and Wikipedia, is in an earlier stage of development; this is especially true for the German edition.

GermaNet–Wiktionary. Unlike OmegaWiki, Wiktionary's coverage of lexical items is almost the same as GermaNet's ($>99\%$; see Table 3), which leads to a higher number of affected items in the test data set and, consequently, significantly better overall results in comparison to OmegaWiki in the same setup. For nouns and verbs, the clustering yields major improvements (Table 2), while the benefit for adjectives is modest. However, it comes as a surprise that the results are not even better—if for almost every lexeme alignment targets can be found, the assumption is that many clusters could be formed. This is not the case as on the test data set, the degree of polysemy is almost the same in both resources, and GermaNet is substantially less polysemous for verbs. Hence, for many senses in GermaNet there exists an equivalent sense with comparable granularity in Wiktionary, and no 1:n mapping can be found which would imply a clustering.

	GN	**OW**	**WKT**	**WP**
Nouns cov. (%)	100.0	20.6	99.9	80.6
Verbs cov. (%)	100.0	20.7	99.9	—
Adjs. cov. (%)	100.0	29.8	98.6	—
Items cov. (%)	100.0	21.4	99.8	45.6
Senses / noun	2.82	1.18	3.84	2.25
Senses / verb	3.70	1.31	3.59	—
Senses / adj.	2.48	1.26	3.24	—
Senses / item	3.21	1.23	3.69	2.25

Table 3: Coverage of lexical items in the test set per resource, and the degree of polysemy (i.e., the average number of senses per item).

While this impairs even better results for our clustering approach, it is also a strong indicator of the quality of the German Wiktionary. Its superiority in certain respects over the English version has already been described by Meyer (2013).

When both approaches are combined, recall is again considerably higher, but the overall results

are not—more items are affected, but no more can be improved (see Table 1). Here, we apparently hit the limits of the clustering approach: While large clusters (and many affected items) are generally desirable, a certain level of precision has to be maintained for this approach to be effective.

GermaNet–Wikipedia. As Wikipedia contains almost exclusively noun concepts, our evaluation for this clustering was restricted to this part of speech (see Table 2). We observe that the results for Dijkstra-WSA alone as well as for the similarity-based approach are significantly better than random, but worse than for the other clusterings. This is explicable by the fact that the polysemy for nouns is comparable for GermaNet and Wikipedia (see Table 3). The observation made for Wiktionary that similar granularity implies many 1:1 alignments and thus few and small clusters holds here as well, as many GermaNet noun senses in the data set have a corresponding entry in Wikipedia. An example is the noun *Filter*, where GermaNet encodes three senses (filter for liquids, air filter, and polarization filter) which are all present in Wikipedia and correctly aligned. Due to its encyclopedic focus, Wikipedia also contains senses which are rather obscure and unlikely to be found in a dictionary (e.g., *Filter* is also an American rock band). Our analysis shows, however, that the alignment algorithm reliably rules them out as alignment targets so that they usually do not impair the clustering outcome.

When combining both approaches in the hybrid setup, we get the expected boost in recall, and the significantly better WSD result (+2.0 as compared to the random setup) suggests that the precision is still acceptable. This is in line with the results reported in (Matuschek and Gurevych, 2013) on the task of WordNet–Wikipedia alignment, which is comparable due to the similar structures of WordNet and GermaNet; in this setup, the hybrid approach yielded better recall while maintaining the same precision as the individual approaches.

Combined approaches. Our experiments show that clustering GermaNet against different collaboratively constructed LSRs using a state-of-the-art WSA algorithm is indeed effective: with few exceptions, the WSD results beat comparable random clusterings, and often significantly so.

A main insight was that different clusterings do not work equally well on each part of speech: while OmegaWiki works best for adjectives, Wiktionary gives the best results for nouns and verbs. Thus, we performed an additional experiment where optimal clusterings were chosen for each part of speech (the boldface results from Table 2). This clustering yields a significant improvement in WSD for each part of speech except adjectives, and achieves the strongest overall improvement (1.9 percentage points) over random clustering. This shows that our language-independent approach is effective, even though it consists solely of an alignment algorithm which does not rely on any resource-specific tuning or knowledge external to any of the resources involved. This is in strong contrast to previous work such as Snow et al. (2007), who employ further external resources, as well as features specifically tailored towards WordNet in a supervised machine learning setup.

4.4 Experiments on WordNet

To demonstrate the validity of our approach for English, we also clustered WordNet by aligning it to the English editions of the three collaboratively constructed LSRs and used the resulting coarse-grained WordNet for WSD. We rescored the raw sense assignments of the three top-performing systems in the Senseval-3 English all-words WSD task (Snyder and Palmer, 2004); the results, averaged across all systems, are shown in Table 4. In general, our observation of significantly improved WSD performance held for English as well. While there are some deviations from the results we reported for German, the observations regarding the properties of the collaboratively constructed LSRs can for the most part be transferred.

As for German, we observed that different clusterings do not work equally well on each part of speech. Thus, we also tested a configuration for English where we selected the optimal clusterings for each part of speech (the boldface results from Table 4). As with German, this clustering results in a significant improvement for each part of speech (except adverbs, though these comprise only 15 of the 2041 instances in the data set).

			OmegaWiki			Wiktionary			Wikipedia		
		none	**rand.**	**WSA**	±	**rand.**	**WSA**	±	**rand.**	**WSA**	±
no backoff	noun	69.0	70.2	71.0	0.8*	70.7	71.4	0.6	71.5	72.5	1.0*
	verb	56.4	59.5	61.2	**1.8***	63.8	64.9	1.1	—	—	—
	adj.	69.3	69.8	69.7	0.0	70.5	70.9	0.5	—	—	—
	adv.	86.7	86.7	86.7	0.0	86.7	86.7	0.0	—	—	—
	all	64.6	66.4	67.4	**1.0***	68.3	69.1	0.8*	—	—	—
w/backoff	noun	69.0	78.4	80.5	**2.2***	72.6	73.6	1.0*	73.5	74.2	0.8
	verb	56.4	69.5	66.9	−2.6*	65.4	66.5	1.0	—	—	—
	adj.	69.3	78.9	82.4	**3.4***	73.6	74.0	0.4	—	—	—
	adv.	86.7	86.7	86.7	0.0	86.7	86.7	0.0	—	—	—
	all	64.6	75.3	76.0	0.7	70.3	71.2	0.9*	—	—	—

Table 4: WSD accuracy (F-score) by POS, using clusterings derived from Dijkstra-WSA alignments of WordNet to various resources, without (top) and with (bottom) the similarity-based backoff. Boldface marks best results per POS; asterisks mark statistically significant differences from the random baseline.

5 Conclusions and future work

In this work, we presented a method for clustering fine-grained GermaNet senses by aligning them to three different collaboratively constructed sense inventories. We used Dijkstra-WSA, a language-independent alignment algorithm which is easily applicable to a variety of LSRs. We showed that a significant improvement in word sense disambiguation accuracy is possible with this method. In contrast to previous approaches, ours is substantially more flexible and generic, relying on no knowledge external to the LSRs and no resource-specific feature engineering. As evidence of this, we demonstrated that our method also performs well with the English WordNet. We also discussed the properties of the different LSRs regarding coverage and granularity, and showed that combining clusterings of different resources for different parts of speech leads to the best performance. Our clusterings will be made freely available to the research community at `https://www.ukp.tu-darmstadt.de/data/`.

One task we intend to investigate in future work is an evaluation on the forthcoming sense-annotated extension to the TüBa-D/Z corpus (Henrich et al., 2013). And as Dijkstra-WSA is applicable to arbitrary pairs of LSRs, we would also like to investigate clustering LSRs other than GermaNet and WordNet, which are by far not the only ones with a tendency towards microdistinction of senses (Jorgensen, 1990). Not only might this improve performance when these sense inventories are used for WSD, but it might also help in the curation of these resources by identifying questionable sense distinctions. This seems especially interesting for Wiktionary and OmegaWiki, which have quite different sense granularities but whose collaborative construction model allow for easy revision of entries.

Regarding improvements to the clustering approach itself, we would like to evaluate to what extent the clusters we create respect the existing taxonomic structure of the resources induced by semantic relations; for instance, merging senses on different levels of the GermaNet taxonomy could lead to circular or otherwise contradictory relations. Following Snow et al. (2007), we want to investigate how such violations of the taxonomy can be avoided in the algorithmic approach.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant № I/82806, and by the Hessian research excellence program *Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE)* as part of the research center "Digital Humanities". The authors thank Verena Henrich for providing raw sense assignments for her systems on WebCAGe 2.0.

References

Eneko Agirre and Oier Lopez de Lacalle. 2003. Clustering WordNet word senses. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 11–18, September.

Sumit Bhagwani, Shrutiranjan Satapathy, and Harish Karnick. 2013. Merging word senses. In *Proceedings of Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-8)*, pages 11–19.

Samuel Broscheit, Anette Frank, Dominic Jehle, Simone Paolo Ponzetto, Danny Rehl, Anja Summa, Klaus Suttner, and Saskia Vola. 2010. Rapid bootstrapping of word sense disambiguation resources for German. In Manfred Pinkal, Ines Rehbein, Sabine Schulte im Walde, and Angelika Storrer, editors, *Proceedings of the 10th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2010)*, pages 19–27, September.

Paul Buitelaar. 2000. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In Amit Bagga, James Pustejovsky, and Wlodek Zadrozny, editors, *Proceedings of the NAACL-ANLP 2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*, pages 14–19, April.

Jen Nan Chen and Jason S. Chang. 1998. Topical clustering of MRD senses based on information retrieval techniques. *Computational Linguistics*, 24(1):61–95.

Timothy Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, September.

Edsger W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

William B. Dolan. 1994. Word sense ambiguation: Clustering related senses. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, volume 2, pages 712–716, August.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Iryna Gurevych and Jungi Kim, editors. 2012. *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Theory and Applications of Natural Language Processing. Springer.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY – A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – A lexical-semantic net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Verena Henrich and Erhard Hinrichs. 2012. A comparative evaluation of word sense disambiguation algorithms for German. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, May.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2011. Semi-automatic extension of GermaNet with sense definitions from Wiktionary. In *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, pages 126–130.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2012. WebCAGe – A Web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 387–396, April.

Verena Henrich, Erhard Hinrichs, and Reinhild Barkey. 2013. Extending the TüBa-D/Z treebank with GermaNet sense annotation. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Proceedings of the 25th Conference of the German Society for Computational Linguistics (GSCL 2013)*, volume 8105 of *Lecture Notes in Artificial Intelligence*, pages 89–96. Springer, September.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27, January. (Introduction to the special issue "Artificial Intelligence, Wikipedia and Semi-Structured Resources").

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*, chapter 3. Springer.

Nancy Ide. 2006. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Alexander Gelbukh, editor, *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing*

2006), volume 3878 of *Lecture Notes in Computer Science*, pages 13–27. Springer, February.

Julia C. Jorgensen. 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19(3):167–190.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.

Andrew Krizhanovsky. 2012. A quantitative analysis of the English lexicon in Wiktionaries and WordNet. *International Journal of Intelligent Information Technologies*, 8(4):13–22.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 1:151–164, May.

Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, pages 17–24.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford University Press.

Christian M. Meyer. 2013. *Wiktionary: The Metalexicographic and the Natural Language Processing Perspective*. Ph.D. thesis, Technische Universität Darmstadt, `http://tuprints.ulb.tu-darmstadt.de/3654/`, October.

Rada Mihalcea and Dan I. Moldovan. 2001. Automatic generation of a coarse grained WordNet. In *Proceedings of the SIGLEX Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Rada Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Procedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, pages 196–203.

Tristan Miller, Nicolai Erbs, Hans-Peter Zorn, Torsten Zesch, and Iryna Gurevych. 2013. DKPro WSD: A generalized UIMA-based framework for word sense disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 37–42, August.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 25–30.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Association for Computational Linguistics (ACL-COLING 2006)*, pages 105–112.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In Robert Porzel, editor, *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, May.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163, June.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic sense clustering in EuroWordNet. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, pages 409–416.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning (EMNLP-CoNLL 2007)*, pages 1005–1014, June.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 41–43, July.

Catherine Soanes and Angus Stevenson, editors. 2003. *Oxford Dictionary of English*. Oxford University Press.

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*.

Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing Wikipedia as a lexical semantic resource. In *Data Structures for Linguistic Resources and Applications*, pages 197–205. Narr, Tübingen, Germany, April.